Building a forensic ancestry panel from the ground up

The EUROFORGEN Global AIM-SNP set

Phillips, C; Parson, W; Lundsberg, Birgitte Møller; Santos, C; Freire-Aradas, A; Torres, M; Eduardoff, M; Børsting, C; Johansen, P; Fondevila, M; Morling, N; Schneider, P; Carracedo, A; Lareu, M V; EUROFORGEN-NoE Consortium

Published in:
Forensic science international. Genetics

DOI:
10.1016/j.fsigen.2014.02.012

Publication date:
2014

Forensic Science International: Genetics 11 (2014) 13–25

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Building a forensic ancestry panel from the ground up: The EUROFORGEN Global AIM-SNP set

C. Phillips [a,*], W. Parson [b,c], B. Lundsberg [d], C. Santos [a], A. Freire-Aradas [a], M. Torres [e], M. Eduardoff [b], C. Børsting [d], P. Johansen [d], M. Fondevila [a], N. Morling [d], P. Schneider [f] the EUROFORGEN-NoE Consortium, Á. Carracedo [a,e,g], M.V. Lareu [a]

[a] Forensic Genetics Unit, Institute of Legal Medicine, Faculty of Medicine, University of Santiago de Compostela, ES-15705 Santiago de Compostela, Galicia, Spain
[b] Institute of Legal Medicine, Innsbruck Medical University, Müllerstrasse 44, A-6020 Innsbruck, Austria
[c] Penn State Eberly College of Science, University Park, PA, USA
[d] Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Frederik V's Vej 11, DK-2100 Copenhagen, Denmark
[e] Grupo de Medicina Xenómica (GMX), Faculty of Medicine, University of Santiago de Compostela, Spain
[f] Institute of Legal Medicine, University Hospital Cologne, D-50823 Cologne, Germany
[g] Center of Excellence in Genomic Medicine Research, King Abdulaziz University, Jeddah, Saudi Arabia

ARTICLE INFO

Article history:
Received 18 July 2013
Received in revised form 12 February 2014
Accepted 17 February 2014

Keywords:
SNPs
AIMs
HGDP-CEPH
1000 Genomes
SPSmart
Forensic ancestry analysis

ABSTRACT

Emerging next-generation sequencing technologies will enable DNA analyses to add pigmentation predictive and ancestry informative (AIM) SNPs to the range of markers detectable from a single PCR test. This prompted us to re-appraise current forensic and genomics AIM-SNPs and from the best sets, to identify the most divergent markers for a five population group differentiation of Africans, Europeans, East Asians, Native Americans and Oceanians by using our own online genome variation browsers. We prioritized careful balancing of population differentiation across the five group comparisons in order to minimize bias when estimating co-ancestry proportions in individuals with admixed ancestries. The differentiation of European from Middle East or South Asian ancestries was not chosen as a characteristic in order to concentrate on introducing Oceanian differentiation for the first time in a forensic AIM set. We describe a complete set of 128 AIM-SNPs that have near identical population-specific divergence across five continentally defined population groups. The full set can be systematically reduced in size, while preserving the most informative markers and the balance of population-specific divergence in at least four groups. We describe subsets of 88, 55, 28, 20 and 12 AIMs, enabling both new and existing SNP genotyping technologies to exploit the best markers identified for forensic ancestry analysis.

1. Introduction

The prospects for typing 200–300 single nucleotide polymorphisms (SNPs) in one multiplexed sequencing analysis are now much more realistic with the emergence of fast, compact next-generation sequencing systems (NGS), such as Life Technologies Ion Torrent and Illumina MiSeq [1,2]. SNPs have the benefit of complimenting conventional forensic STR analysis by providing information about the DNA donor that can progress a criminal investigation lacking any leads beyond knowledge of gender. Principal amongst the complimentary data generated by SNP analysis is the inference of genetic ancestry and prediction of common physical traits, with SNP-based analysis of pigmentation now established as a viable investigative tool [3–5]. Until the development of compact NGS approaches, forensic ancestry analysis centered on small-scale multiplexes of carefully chosen SNPs and Indels, exemplified by a 34-SNP SNaPshot multiplex and a 46-Indel dye-labeled PCR multiplex [6–8]. Once optimized, we successfully applied these tests to a variety of challenging DNA cases [9–12] and their combination into 80-marker profiles provides good data depth, short-amplicon PCRs sensitive to degraded DNA and complimentary features including Indel's enhanced ability to detect mixed DNA. However, the original choice of ancestry informative markers, particularly components

* Corresponding author. Tel.: +34 981 583 015.
E-mail address: c.phillips@mac.com (C. Phillips).

http://dx.doi.org/10.1016/j.fsigen.2014.02.012
1872-4973/

of the 34-plex SNP test, reflected the state of knowledge of human SNP variation some nine years ago. Now much more extensive SNP catalogs can be screened for suitable candidate markers with major human genome initiatives including HapMap, 1000 Genomes and Complete Genomics publicly releasing project data to allow identification of the best markers for ancestry inference purposes.

We decided to build, from a completely refreshed list of candidates, a new ancestry SNP (AIM-SNP) panel using our own bio-informatics search tools [13,14] that front-end public genome data. Reconfiguring a forensic AIM-SNP set allows several characteristics to be prioritized: (i) identifying the most powerful differentiators for each population comparison; (ii) finding alternative loci with near-identical frequency distributions due to LD-block correlations [15] when SNP multiplexing problems arise, and (iii) carefully balancing marker combinations to give equivalent levels of differentiation between population groups comprising: Africans, Europeans, East Asians, Native Americans and Oceanians. The third characteristic is the most desirable for ensuring less biased assessments of admixture proportions in individuals with detectable co-ancestry–a significant demographic feature of urban populations and regions with histories of population movement (see Chapter 14 of [16]). However, population differentiation balance is also the most challenging characteristic to achieve, since, of the above five groups, Native American and Oceanian variation is not represented in any of the full human SNP catalogs. Luckily, more than 650,000 SNPs have been characterized for the CEPH Human Genome Diversity Panel (HGDP-CEPH) with two Oceanian populations and five American populations [17], so suitable SNPs can be identified for differentiating these two groups, albeit from much smaller sample sizes.

This paper outlines the AIM-SNPs chosen to construct a set of 128 markers suitable for inclusion in forensic NGS tests. The set maintains near-identical population differentiation balance between admixture contributors originating from the five main continentally defined population groups. Therefore the AIM-SNPs together allow analysis of admixed individuals, provided the co-ancestry contributors themselves are not admixed. The AIMs are applicable to a large proportion of the worldwide distribution of human populations, including regions where populations meet and admixture contributors are not necessarily confined to Europeans, Africans or East Asians, e.g. American contributors in the USA and South America or Oceanians in Australia. However, differentiation of European from Middle East or South Asian sub-groups of Eurasia was ignored in favor of ensuring Oceanian differentiation comparable to the other groups. The possibility of allele frequency bias in the populations used to select AIM-SNPs can still exist so we attempted to minimize this by using at least two geographically separated populations per group. Four populations likely to be divergent from those used for selection were also tested to gauge the degree of allelic heterogeneity they exhibited for the same SNPs. Because size constraints can still apply to PCR multiplexes in all technologies, (forensic NGS tests may include STRs as main components), we also reduced the SNP set to smaller scale subsets while maintaining the population differentiation balance at each stage of reduction. Lastly, we describe Sequenom iPLEX® MALDI-TOF genotyping tests used to validate additional population variation in the AIM-SNPs chosen and to assess each SNP's multiplex performance ahead of porting them to larger-scale NGS chemistries.

2. Materials and methods

2.1. Sources of AIM-SNPs and allele frequencies in the five main global population groups

Candidate AIM-SNPs were compiled from three sources: (i) SNP sets previously developed for a range of forensic ancestry test initiatives at Santiago (USC); (ii) allele frequency screens of the Stanford HGDP-CEPH 650 K SNP dataset [17,18] – identifying SNPs with the highest divergence between targeted population comparisons by finding the top 5% most differentiated in each case, and (iii) AIM-SNP lists published both before and after availability of whole genome scan (WGS) high-density SNP arrays. This third approach collected a large number of SNPs highly informative for ancestry but previously systematically excluded from WGS SNP sets due to their lack of association power (e.g. rs16891982 shows a highly differentiated G allele, close to fixation in Europeans and therefore uninformative for correlated variation in SLC45A2 where it is sited). Mainly, SNPs identified as good AIMs prior to use of WGS arrays are commonly part of forensic ancestry sets and include many of the best population differentiators such as rs2814778, rs16891982, rs1426654 and rs3827760, all absent from WGS sets. As Stanford HGDP-CEPH SNP data uses Illumina 650 K WGS array genotyping, the second strategy appears contradictory. However, we identified Oceanian and Native American informative SNPs from Stanford HGDP-CEPH data, as Illumina selected the 650,000 loci based solely on European, African and East Asian variability. Sources of SNP variation we scrutinized are outlined in Fig. 1. Published SNP sets were: 178 of the *DNAprint Ancestry-by-DNA* forensic set of Halder et al. [19]; 128 of the American population analysis panel of Kosoy et al. [20,21]; 445 of the Latin American Cancer Epidemiology (LACE) group's American population analysis panel of Galanter et al. [22]; 47 of the forensic ancestry panel of Kersbergen et al. [23], and; 55 listed in FROGkb, forming several Kiddlab forensic ancestry panels [24]. AIM-SNPs from previous USC developments comprised: 34 of an established forensic ancestry panel [7,10]; 28 of a European-African admixture detection panel (unpublished); 25 of a CT-SNP set developed for enhanced SNaPshot peak balance (unpublished); 46 of a USC-Applied Biosystems Genplex forensic AIM-SNP set, developed in 2007–8 but not released as a kit [25], and 48 of a Sequenom iPLEX® case-control association study (CCAS) stratification adjustment panel [26]. A further ~260 American- and Oceanian-informative SNPs were identified by screening Stanford HGDP-CEPH data, including 28 SNPs already forming a dedicated Oceanian-informative forensic SNaPshot multiplex, termed *Pacifiplex* (publication in preparation). Nine Oceanian- and six East Asian-informative SNPs were novel selections for this study. The above ten sets provided 1031 SNPs for scrutiny. There was some SNP commonality between USC and independently published forensic ancestry sets, comprising: rs16891982, rs2814778 in Kiddlab's FROGkb set and *DNAprint*; rs12913832, rs1426654, rs2471552, rs2715883 in the FROGkb set (only rs12498138 in the published Kiddlab set [24]) and rs3827760 in Kersbergen's set.

Candidate SNP allele frequencies were collected for African (herein AFR), European (EUR) and East Asian (E ASN) reference populations in 1000 Genomes using the USC *ENGINES* online browser [14]. The *ENGINES* portal accesses ~28 million variant sites in 14 populations but we chose the ten unadmixed populations combined in three groups to assess candidate's allele frequency distributions. Frequencies were compiled from: 97 Luhya in Webuye, Kenya (LWK) and 88 Yoruba in Ibadan, Nigeria (YRI) combined as 185 AFR; 87 Utah residents with north and west European ancestry from the CEPH collection (CEU), 93 Finnish in Finland (FIN), 88 British in England and Scotland (GBR), 98 Tuscans in Italy (TSI), 14 Iberians from Spain (IBS) = 380 EUR; 89 Japanese in Tokyo, Japan (JPT), 97 Han Chinese in Beijing, China (CHB), 100 Han Southern Chinese (CHS) = 286 E ASN. 1000 Genomes data of four admixed populations: 61 individuals of African ancestry in Southwest USA (ASW), 66 individuals of Mexican ancestry in Los Angeles, California (MXL), 55 Puerto Ricans in Puerto Rico (PUR) and 60 Colombians in Medellín (CLM) was accessed later to assess the ability of selected AIM-SNPs to gauge admixture.

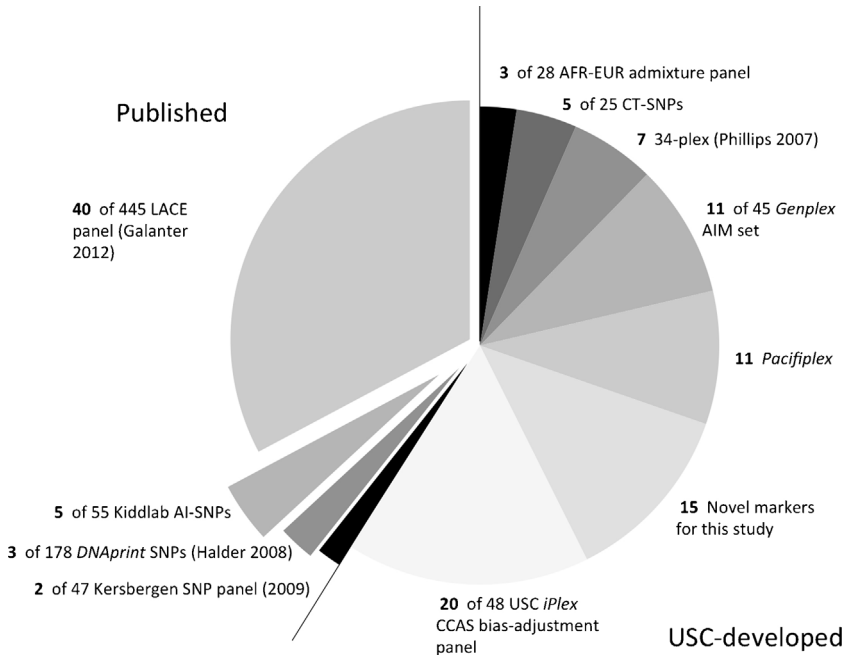

Fig. 1. Sources of AIM-SNP candidates and numbers selected from each source.

American (herein AME) and Oceanian (OCE) allele frequencies were collected from Stanford HGDP-CEPH SNP data using the USC SPSmart online browser [13] to access 14 Brazilian Karitiana, 8 Brazilian Surui, 7 Colombian Piapoco, 21 Mexican Maya, 14 Mexican Pima = 64 AME plus 11 Bougainville Melanesian, 17 Papua New Guinean = 28 OCE.

2.2. Selection criteria for the final AIM-SNP set

Relative population differentiation power for each candidate was estimated by ranking SNPs by allele frequency differential (or delta: δ), given by the absolute value of the frequency of an allele in one population minus the frequency in another. Pairwise population comparison deltas were summed for the three main groups, e.g. AFR = (AFR-EUR δ) + (AFR-E ASN δ). Delta correlates closely to the other widely used population differentiation metrics: Fst and Rosenberg's $\mathbf{I_n}$ Divergence [27], e.g. Fst$\approx\delta/(2-\delta)$ (and see Fig. 2 of [28]). Fst and $\mathbf{I_n}$ metrics were applied once markers had been ordered by their population-specific differentiation values, but note Fst and $\mathbf{I_n}$ normally measure overall population divergence, whereas summed δ estimates provide the optimum way to identify best markers for any one pairwise population differentiation. This is particularly important when selecting SNPs to differentiate closely related populations that share recent demographic history and therefore show reduced divergence, as demonstrated by most E ASN-AME allele frequency differences. OCE-informative SNPs had not been identified in any previously compiled AIM-SNP sets so we accessed USC lists of the top 5% most divergent markers compiled to construct *Pacifiplex*.

A minimum one megabase (Mb) inter-marker spacing filter was applied to prevent multiple, highly divergent SNPs from the same genomic segment (i.e. in linkage disequilibrium: LD) occupying the top of each population differentiation ranking. Choosing multiple markers with limited recombination between them can bias co-ancestry proportion estimates in admixed individuals. Furthermore, LD can sometimes extend beyond 1 Mb – notably in the very high extended haplotype homozygosities seen in the LCT region due to rapid recent evolution in North Europe [29]. However, such instances are well-documented, so we checked LD using HapMap D' estimates between only the most closely sited SNP pairs. The same phenomenon of allele frequency correlations in closely sited SNPs was exploited to broaden the choice of alternative candidates when first-strike markers failed in multiplex. We previously observed that a series of SNPs with identical allele frequencies occurred in succession in a large number of locations in the human genome [15]. Therefore alternative markers are often available to select with no change in their differentiation power. We used the SPSmart chromosome segment query option when alternative SNPs were required and one example in the VRK1 gene is shown in Supplementary Fig. S1A.

A key goal was balancing the degree of differentiation as equally as possible amongst all five population groups. To achieve this we calculated the population-specific Divergence (PSD) $\mathbf{I_n}$ value in each population group for each SNP candidate. PSD is equivalent to the locus specific branch length (LBSL) $\mathbf{I_n}$ measurement described by Galanter et al. and used to balance AFR-EUR-AME population divergence in the LACE SNP set [22]. PSD values were calculated by uploading 1000 Genomes and Stanford HGDP-CEPH SNP genotypes into the *Snipper* forensic ancestry analysis portal [7,11] – deriving each SNP's PSD by pairwise labeling of uploaded populations: i.e. AFR:non-AFR (all other groups), EUR:non-EUR, etc. *Snipper* calculates Divergences from uploaded SNP data, but values are higher than Rosenberg's $\mathbf{I_n}$ ($\mathbf{I_n}$ = *Snipper* Divergence × 0.69), with cumulative values obtained by addition. We first identified the most powerful OCE – then AME-informative SNPs followed by those of the other three groups, carefully gauging the cumulative values until reaching a point of PSD convergence while aiming for ~120–140 SNPs in combination.

2.3. Marker subsets for smaller-scale multiplexed forensic tests

To aid development of ancestry panels for established forensic technologies such as SNaPshot, the completed set of balanced divergence AIM-SNPs (herein the Global AIM-SNP set) was systematically reduced in size to a minimum number of markers capable of successfully classifying AFR-EUR-E ASN, these plus AME, then these plus OCE, while keeping the PSD values balanced throughout each stage of reduction. Additionally, since forensic

NGS multiplex designs may opt to combine core STRs, identification SNPs and pigmentation-predictive SNPs alongside AIM-SNPs, we also compiled SNP sets suitable for such purposes into sets aiming for ~90 and 50–60 components.

2.4. Validation of candidate and selected AIM-SNPs using Sequenom iPLEX® genotyping

Sequenom iPLEX® MALDI-TOF genotyping was used as the system of choice to: (i) assess multiplex performance of component SNPs, albeit in smaller-scale PCR reactions; (ii) genotype key candidate SNPs with the HGDP-CEPH panel when this data was not available from the Stanford study; (iii) check genotyping concordance between NGS chemistries and an established medium-throughput SNP typing system. A total of 126 autosomal AIM-SNPs plus X- and Y-chromosome SNPs were amplified in 5 multiplexes (termed Global 1 to 5) using the Sequenom iPLEX® system as detailed in Supplementary File S1. All iPLEX® primer sequences are listed in Supplementary Table 1A.

The sensitivity of the Global 1-5 iPLEX tests was examined twice in two samples at dilutions: 20, 10, 5,1.25 0.5 ng/μL. DNA concentrations were determined using the Qubit® dsDNA BR Assay Kit on the Qubit® 2.0 Fluorometer (LT), following manufacturer's guidelines.

Sanger sequencing analyzed SNPs: rs2789823; rs8137373; rs1557553; rs2471552; rs12913832, showing reproducible peak imbalances, to check for uncharted primer binding site SNPs or Indels. Flanking regions were amplified by PCR using AmpliTaq Gold® DNA polymerase (LT) in GeneAmp® 10X PCR Buffer with 10 mM dNTP in 25 μL volumes (primer sequences in Supplementary Table S1B). PCR conditions were: 3 min at 94 °C then 35 cycles of 30 s at 94 °C, 30 s at 60 °C, 30 s at 72 °C then 72 °C for 10 min. PCR products were purified with 4 μL Exo-SAP IT (Affymetrix) to 10 μL PCR products, incubating for 15 min at 37 °C then 15 min at 80 °C. PCR products were sequenced using the BigDye Terminator Kit (LT) following manufacturer's recommended protocols. Sequences were analyzed on a 3130xl Genetic Analyzer (LT) with 36 cm capillary arrays and POP-4 polymer (LT).

2.5. Tri-allelic SNPs included in the selection process

We previously identified several tri-allelic SNPs exhibiting highly skewed allele frequency distributions amongst the five groups. We included SNPs: rs4540055 and rs5030240, already part of the 34-plex test that successfully identified mixed source DNA from donors with different ancestries (notably when third alleles show contrasting common and rare frequencies in contributor populations [8]). Unlike WGS techniques, sequencing or iPLEX and SNaPshot SBE tests reliably detect three alleles at substitution sites, so another four tri-allelic SNPs were added and their HGDP-CEPH allele frequencies are summarized in Supplementary Fig. S2.

2.6. Population analyses

The Global AIM-SNP set was assessed for ability to infer co-ancestry proportions in 1000 Genomes ASW, MXL, PUR and CLM admixed populations (described in Section 2.1). We compared co-ancestry components detected using 128 Global AIM-SNPs and 350 of 445 LACE SNPs (AME data not available for 95): comprising a large-scale CCAS analysis panel [22] designed to differentiate AFR-EUR-AME admixture components, with E ASN and OCE components not expected to be detectable parts of the above four 1000 Genomes admixed population's ancestry profiles.

We applied standard population analysis approaches of STRUCTURE and Principal Component Analysis (PCA). STRUCTURE (v.2.3.3) was used with CLUMPP and distruct software [31], applying parameters: Burnin = 100,000; MCMC steps = 100,000; Admixture/POPFLAG model with correlated allele frequencies and 5 iterations. PCA used R v.2.11.1 and the *SNPassoc* package [32]. Values of K clusters from 2 to 8 were explored but K:5 was identified as the optimum number of populations from comparison of posterior probabilities given by X|K. Although *Snipper* analysis can also be applied to these assessments, it is qualitative rather than quantitative, i.e. providing a classification based on the two highest ancestry likelihoods but not currently accounting for admixture. However, Supplementary File S2 provides the five group-128 SNP *Snipper* training set, adaptable for any marker subsets by removing columns.

A proportion of populations in any one group can be divergent from those we used for AIM-SNP selection and this may reduce the ability of the marker set to be optimally informative for all regions of a population group's geographic distribution. We assessed within-group variation in SNP allele frequencies by analyzing 62 Greenlanders, 32 Fijians and 77 Somalis in 116/128 SNPs with these additional samples exemplifying geographically distinct populations in America, Oceania and Africa. We also typed 22 Pathan Pakistanis from the HGDP-CEPH panel to compare to European allele frequency distributions. In addition, within-group variation of SNPs reported for HGDP-CEPH by Stanford were analyzed, although this data set comprises 75/128 marker coverage. SNP data from the 45 extra populations was analyzed using Arlequin (v.3.5.1.3) and STRUCTURE.

3. Results

3.1. Characteristics of the ancestry informative SNPs selected

A final set of 122 bi-allelic and 6 tri-allelic SNPs were selected from a total candidate pool of 189 loci (and 12 tri-allelic) and are detailed in Table 1 and Supplementary Table S2A. All candidate SNPs from sources detailed in Section 2.1 are listed in Supplementary Table S2B. Global AIM-SNP allele frequency distributions in five population groups are summarized in Fig. 2. The cumulative PSD values in each group required a smaller number of AFR-informative SNPs and for 28 candidates, Oceanian and American allele frequency data was absent from the Stanford HGDP-CEPH study so these loci were kept in reserve. This pool of alternative SNPs able to substitute loci failing in multiplex comprised: 6 tri-allelics; 12 AFR-; 17 EUR-; 7 AME-; 2 E ASN- and one OCE-informative SNP (rows 138–203, Supplementary Table S2A).

To measure the overall population differentiation power of the 128 Global AIM-SNPs, we compared their individual Fst values for three, four and five group comparisons with SNPs in the Kosoy ancestry set [20]. The Kosoy set also has 128 AIM SNPs and only rs9522149 in common. Observed Fst values are shown in ascending order in Fig. 3. Average Fst values for AFR-EUR-E ASN comparisons are Kosoy = 0.260, Global = 0.378, AFR-EUR-E ASN-AME comparisons: 0.321 vs. 0.438 and AFR-EUR-E ASN-AME-OCE: 0.309 vs. 0.487. Overall, only 38 of 384 SNP comparisons indicate higher differentiation power for Kosoy SNPs, none for Fst values for AFR-EUR-E ASN-AME-OCE comparisons. As with the LACE panel, Kosoy's SNP set aims to primarily differentiate AFR, EUR and AME co-ancestry components as these are found in a large proportion of US and Latin American admixed individuals used in association studies. Therefore many individual Fst values should be higher for AFR-EUR-E ASN-AME comparisons. However, only 21 Kosoy SNPs have higher Fst values for this comparison. The largest contrast of average Fst is observed in AFR-EUR-E ASN-AME-OCE comparisons, with 63% higher differentiation in Global AIM-SNPs. Lastly, a noticeable jump in Fst values is seen in the four and five-group charts for about 10 Global AIM-SNPs at the far right. These could be considered the core forensic AIM-SNP set, showing the highest

Table 1
Details of 128 markers selected for the Global AIM-SNP set.

AIM details							Reference allele frequencies					Individual population-specific divergence (PSD) and cumulative PSD (CD) values									
PG	SNP ID	C	Source	Position	Gene	RA	Afr	Eur	Asn	Ame	Oce	Afr PSD	Afr CD	Eur PSD	EurCD	AsnPSD	AsnCD	Ame PSD	Ame CD	OCE PSD	OCE CD
1	rs9908046	17	USC CCAS panel	53563782	–	C	0.965	0.949	0.879	0.992	0.018	0.010	0.01	0.008	0.01	0.002	0.00	0.024	0.02	0.522	0.52
2	rs2139931	1	USC CCAS panel	84590527	PRKACB	A	0.897	0.774	0.904	0.898	0.018	0.008	0.02	0.006	0.01	0.013	0.01	0.007	0.03	0.424	0.95
3	rs715605	22	Novel	30640308	–	T	0.851	0.913	0.991	1	0.089	0.006	0.02	0.000	0.01	0.034	0.05	0.025	0.06	0.415	1.36
4	rs3751050	11	Pacifiplex	9091244	SCUBE2	A	0.949	0.888	0.96	0.961	0.089	0.006	0.03	0.001	0.01	0.012	0.06	0.008	0.06	0.411	1.77
5	rs6054465	20	USC CCAS panel	6673018	–	T	0.962	0.87	0.748	0.859	0.036	0.036	0.07	0.005	0.02	0.010	0.07	0.001	0.06	0.396	2.17
6	rs26951	5	Novel	59759657	PDE4D	G	0.997	0.907	0.629	0.82	0.036	0.083	0.15	0.024	0.04	0.047	0.12	0.000	0.06	0.380	2.55
7	rs6886019	5	USC CCAS panel	170245846	–	C	0.941	0.942	1	0.883	0.161	0.000	0.15	0.001	0.04	0.031	0.15	0.004	0.07	0.374	2.92
8	rs10970986	9	USC CCAS panel	32453278	–	T	0.981	0.726	0.734	0.711	0.018	0.085	0.24	0.002	0.05	0.001	0.15	0.002	0.07	0.362	3.28
9	rs16830500	2	Novel	152814129	CACNB4	T	0.938	0.939	0.428	0.875	0	0.049	0.28	0.079	0.12	0.128	0.28	0.014	0.08	0.359	3.64
10	rs3804030	21	Novel	45629565	–	A	0.811	0.903	0.865	0.992	0.089	0.003	0.29	0.001	0.13	0.005	0.28	0.043	0.13	0.356	4.00
11	rs2274636	10	USC CCAS panel	27443012	MASTL	A	0.995	0.882	0.719	0.906	0.107	0.066	0.35	0.006	0.13	0.022	0.30	0.007	0.13	0.314	4.31
12	rs4806654	19	Novel	55768276	PPP6R1	T	0.719	0.924	0.886	0.813	0.125	0.018	0.37	0.020	0.15	0.004	0.31	0.001	0.14	0.303	4.62
13	rs4391951	13	Novel	44755071	–	T	0.892	0.841	0.477	0.883	0.036	0.035	0.41	0.027	0.18	0.069	0.38	0.024	0.16	0.303	4.92
14	rs10149275	14	Novel	43268640	–	G	0.881	0.903	0.26	0.805	0.018	0.033	0.44	0.097	0.28	0.188	0.56	0.013	0.17	0.295	5.21
15	rs1509524	4	USC CCAS panel	125455038	–	A	0.784	0.795	0.82	0.844	0.089	0.000	0.44	0.000	0.28	0.002	0.57	0.004	0.18	0.288	5.50
16	rs7623065	3	Pacifiplex	22385375	–	A	0.251	0.684	0.871	0.969	0	0.131	0.57	0.001	0.28	0.061	0.63	0.101	0.28	0.284	5.79
17	rs10811102	9	Pacifiplex	1911291	–	G	0.995	0.774	0.32	0.336	0.018	0.181	0.75	0.033	0.31	0.101	0.73	0.049	0.33	0.265	6.05
18	rs10455681	6	Pacifiplex	69802502	BAI3	A	0.989	0.846	0.178	0.445	0.018	0.175	0.93	0.083	0.39	0.216	0.94	0.018	0.34	0.260	6.31
19	rs7832008	8	Pacifiplex	98358246	–	G	0.53	0.137	0.612	0.156	1	0.016	0.94	0.100	0.49	0.054	1.00	0.038	0.38	0.257	6.57
20	rs9809818	3	Pacifiplex	71480566	FOXP1	C	0.019	0.088	0.871	0.82	0.982	0.167	1.11	0.152	0.65	0.264	1.26	0.116	0.50	0.254	6.82
21	rs9934011	16	Novel	13915807	–	T	0.814	0.816	0.217	0.68	0.018	0.041	1.15	0.072	0.72	0.160	1.42	0.004	0.50	0.247	7.07
22	rs3784651	15	Pacifiplex	94925273	MCTP2	T	0.516	0.879	0.248	0.836	0	0.004	1.16	0.137	0.86	0.123	1.55	0.043	0.55	0.240	7.31
23	rs2282107	21	Novel	37707581	MORC3	A	0.881	0.932	0.579	0.867	0.161	0.012	1.17	0.030	0.89	0.040	1.59	0.006	0.55	0.226	7.53
24	rs12405776	1	Pacifiplex	242431557	PLD5	C	0.497	0.979	0.413	0.672	0.089	0.023	1.19	0.200	1.09	0.070	1.66	0.000	0.55	0.203	7.74
25	rs1592672	12	Pacifiplex	80128593	–	T	0.065	0.891	0.248	0.602	0	0.174	1.36	0.253	1.34	0.062	1.72	0.007	0.56	0.182	7.92
26	rs10183022	2	Pacifiplex	237481969	CXCR7	G	0.027	0.611	0.631	0.813	0.982	0.252	1.62	0.010	1.35	0.011	1.73	0.053	0.61	0.173	8.09
27	rs1877751	20	LACE	57967906	–	A	0.849	0.692	0.166	0.031	0.036	0.106	1.72	0.053	1.40	0.124	1.85	0.179	0.79	0.163	8.26
28	rs798949	7	Pacifiplex	120765954	C7orf58	T	0.778	0.525	0.15	0.07	0.929	0.092	1.81	0.010	1.41	0.099	1.95	0.110	0.90	0.158	8.41
1	rs1557553	22	LACE	44760984	–	C	0.962	0.908	0.717	0.086	0.786	0.046	1.86	0.028	1.44	0.011	1.96	0.473	1.37	0.000	8.41
2	rs2080161	7	LACE	13331150	–	T	0.968	0.789	0.713	0.099	0.82	0.075	1.93	0.004	1.44	0.002	1.97	0.415	1.79	0.023	8.44
3	rs10483251	14	LACE	21671277	–	G	0.889	0.779	0.923	0.099	0.712	0.013	1.95	0.000	1.44	0.033	2.00	0.413	2.20	0.005	8.44
4	rs12498138	3	CT-plex	121459589	GOLGB1	G	0.995	0.916	0.914	0.094	0.911	0.046	1.99	0.006	1.45	0.004	2.00	0.411	2.61	0.002	8.44
5	rs8137373	22	USC CCAS panel	41729216	ZC3H7B	G	0.827	0.742	0.923	0.023	0.982	0.004	2.00	0.002	1.45	0.041	2.04	0.394	3.01	0.062	8.51
6	rs1452501	16	LACE	80623262	–	C	0.978	0.959	0.874	0.214	0.635	0.030	2.03	0.028	1.48	0.000	2.04	0.376	3.38	0.042	8.55
7	rs17130385	10	USC CCAS panel	115196019	–	G	0.976	0.934	0.769	0.086	0.518	0.049	2.08	0.034	1.51	0.005	2.05	0.357	3.74	0.058	8.60
8	rs2471552	7	CT-plex	45977173	–	C	0.081	0.211	0.217	0.945	0.107	0.032	2.11	0.001	1.51	0.000	2.05	0.340	4.08	0.015	8.62
9	rs5757362	22	LACE	39306080	–	T	0.014	0.375	0.219	0.935	0.212	0.118	2.23	0.012	1.53	0.007	2.06	0.322	4.40	0.004	8.62
10	rs17359176	13	LACE	23667334	–	G	0.997	0.911	0.822	0.204	0.982	0.062	2.29	0.010	1.54	0.002	2.06	0.320	4.72	0.033	8.66
11	rs174570	11	LACE	61597212	FADS2	C	0.992	0.841	0.593	0.031	0.482	0.115	2.40	0.023	1.56	0.023	2.08	0.319	5.04	0.035	8.69
12	rs10012227	4	USC CCAS panel	18637315	–	G	0.905	0.867	0.528	0.047	0.768	0.044	2.45	0.044	1.60	0.040	2.12	0.304	5.34	0.002	8.69
13	rs11960137	5	LACE	155338081	–	C	0.938	0.859	0.785	0.122	0.964	0.029	2.48	0.007	1.61	0.001	2.12	0.304	5.65	0.035	8.73
14	rs2051827	12	USC CCAS panel	47956031	–	G	0.949	0.932	0.937	0.195	0.518	0.013	2.49	0.011	1.62	0.011	2.13	0.303	5.95	0.085	8.81
15	rs12402499	1	CT-plex	101528954	–	G	1	0.909	1	0.258	1	0.033	2.52	0.000	1.62	0.041	2.17	0.300	6.25	0.014	8.83
16	rs11625446	14	LACE	48244558	–	C	0.951	0.739	0.614	0.043	0.982	0.081	2.60	0.002	1.62	0.009	2.18	0.296	6.55	0.089	8.92
17	rs4979274	9	USC CCAS panel	116444269	–	C	0.995	0.88	0.57	0.086	0.946	0.105	2.71	0.032	1.66	0.044	2.23	0.293	6.84	0.040	8.96
18	rs7151991	14	LACE	32635572	–	A	0.154	0.199	0.15	0.935	0	0.005	2.71	0.001	1.66	0.007	2.23	0.277	7.12	0.057	9.01
19	rs4780476	16	USC CCAS panel	12862007	CPPED1	A	0.227	0.308	0.247	0.945	0.232	0.007	2.72	0.000	1.66	0.006	2.24	0.272	7.39	0.004	9.02
20	rs6088466	20	LACE	32913534	–	G	0.978	0.679	0.572	0.034	0.75	0.131	2.85	0.000	1.66	0.010	2.25	0.266	7.65	0.005	9.02
21	rs2302013	2	USC CCAS panel	242042331	FARP2	T	1	0.997	0.845	0.344	0.339	0.045	2.90	0.057	1.71	0.002	2.25	0.219	7.87	0.168	9.19
22	rs4792928	17	USC CCAS panel	42105174	–	T	1	0.955	0.369	0.195	0.804	0.126	3.02	0.111	1.82	0.154	2.40	0.175	8.05	0.004	9.20
1	rs2814778	1	34plex	159174683	DARC	A	0.003	0.997	1	0.992	1	0.674	3.70	0.123	1.95	0.105	2.51	0.065	8.11	0.051	9.25
2	rs2789823	9	CT-plex	136769888	VAV2	G	0.908	0.003	0	0	0	0.528	4.23	0.109	2.06	0.093	2.60	0.057	8.17	0.044	9.29
3	rs1871534	8	Genplex	145639681	SLC39A4	C	0.886	0.003	0	0	0	0.503	4.73	0.106	2.16	0.091	2.69	0.056	8.23	0.042	9.33
4	rs6875659	5	LACE	175158653	–	G	0.046	0.916	0.956	0.977	1	0.483	5.21	0.051	2.21	0.071	2.76	0.062	8.29	0.065	9.40
5	rs1369290	18	Kersbergen	67691520	RTTN	A	0.881	0.032	0	0	0	0.463	5.67	0.069	2.28	0.097	2.86	0.060	8.35	0.046	9.44

Table 1 (*Continued*)

AIM details							Reference allele frequencies					Individual population-specific divergence (PSD) and cumulative PSD (CD) values									
PG	SNP ID	C	Source	Position	Gene	RA	Afr	Eur	Asn	Ame	Oce	Afr PSD	Afr CD	Eur PSD	EurCD	AsnPSD	AsnCD	Ame PSD	Ame CD	OCE PSD	OCE CD
6	rs310644	20	Kiddlab	62159504	PTK6	A	0.043	0.928	0.965	0.953	0.107	0.454	6.13	0.073	2.36	0.091	2.95	0.051	8.40	0.239	9.68
7	rs1197062	17	LACE	58641118	–	A	0.111	0.953	0.974	1	0.857	0.432	6.56	0.065	2.42	0.072	3.02	0.068	8.47	0.004	9.69
8	rs6034866	20	USC CCAS panel	17603728	RRBP1	A	0.924	0.058	0.082	0.054	0.179	0.430	6.99	0.075	2.50	0.042	3.07	0.045	8.51	0.003	9.69
1	rs1426654	15	34plex	48426484	SLC24A5	A	0.043	0.996	0.014	0.039	0	0.152	7.14	0.617	3.11	0.237	3.30	0.126	8.64	0.145	9.83
2	rs16891982	5	34plex	33951693	SLC45A2	C	0.992	0.029	0.983	0.984	1	0.115	7.26	0.547	3.66	0.229	3.53	0.154	8.79	0.154	9.99
3	rs8072587	17	USC CCAS panel	19211073	EPN2	C	0.978	0.192	1	0.796	1	0.135	7.39	0.370	4.03	0.207	3.74	0.019	8.81	0.122	10.11
4	rs12142199	1	AFR-EUR admix SNPs	1249187	CPSF3L	G	0.954	0.2	0.983	1	1	0.104	7.50	0.369	4.40	0.168	3.91	0.128	8.94	0.107	10.22
5	rs9522149	13	CT-plex	111827167	ARHGEF7	T	0.954	0.258	0.997	0.945	1	0.090	7.59	0.334	4.73	0.175	4.08	0.121	9.06	0.107	10.32
6	rs7531501	1	USC CCAS panel	234338303	–	G	0.065	0.886	0.128	0.258	0.089	0.136	7.72	0.330	5.06	0.109	4.19	0.018	9.08	0.084	10.41
7	rs12913832	15	34plex	28365618	HERC2	A	1	0.289	0.998	0.883	1	0.140	7.86	0.315	5.38	0.169	4.36	0.028	9.11	0.089	10.50
8	rs7084970	10	AFR-EUR admix SNPs	119750413	–	C	0.854	0.063	0.769	0.833	0.214	0.117	7.98	0.294	5.67	0.086	4.45	0.068	9.18	0.044	10.54
9	rs11778591	8	LACE	12720349	–	C	0.624	0.088	0.895	0.786	0.857	0.010	7.99	0.282	5.95	0.175	4.62	0.059	9.23	0.076	10.62
10	rs4749305	10	LACE	28391596	MPP7	A	0.332	0.849	0.051	0.018	0.036	0.007	8.00	0.279	6.23	0.186	4.81	0.173	9.41	0.126	10.74
11	rs820371	3	LACE	123404711	MYLK	A	0.965	0.237	0.949	0.734	0.981	0.123	8.12	0.278	6.51	0.132	4.94	0.003	9.41	0.113	10.86
12	rs1924381	13	LACE	72321856	DACH1	A	0.924	0.13	0.811	0.661	0.786	0.138	8.26	0.272	6.78	0.077	5.02	0.012	9.42	0.034	10.89
13	rs2715883	11	Genplex	120133494	POU2F3	A	0.122	0.708	0.005	0	0	0.039	8.30	0.270	7.05	0.172	5.19	0.115	9.54	0.096	10.99
14	rs595961	1	DNAprint	36367780	EIF2C1	A	0.084	0.859	0.189	0.172	0.464	0.128	8.42	0.264	7.32	0.075	5.26	0.051	9.59	0.000	10.99
15	rs3759171	12	LACE	72307616	TBC1D15	A	0.032	0.743	0.103	0.048	0.018	0.121	8.55	0.264	7.58	0.077	5.34	0.084	9.67	0.109	11.09
16	rs917115	7	Kiddlab	28172586	JAZF1	T	0.178	0.779	0.024	0.156	0.179	0.035	8.58	0.262	7.84	0.184	5.52	0.037	9.71	0.029	11.12
17	rs1486341	12	LACE	39042063	–	A	0.957	0.126	0.733	0.679	0.911	0.183	8.76	0.259	8.10	0.042	5.57	0.019	9.73	0.098	11.22
18	rs11074130	15	LACE	93583742	–	T	0.776	0.096	0.773	0.780	0.804	0.063	8.83	0.257	8.36	0.081	5.65	0.036	9.76	0.055	11.28
19	rs7937598	11	LACE	44745048	–	A	0.778	0.151	0.827	0.982	0.929	0.041	8.87	0.255	8.61	0.083	5.73	0.169	9.93	0.102	11.38
20	rs10186877	2	LACE	216618818	–	G	0.941	0.279	0.902	0.982	0.911	0.090	8.96	0.242	8.86	0.078	5.81	0.127	10.06	0.050	11.43
21	rs1567803	2	LACE	101343018	–	C	0.965	0.322	0.916	0.969	1	0.100	9.06	0.232	9.09	0.074	5.88	0.084	10.14	0.095	11.52
22	rs2889678	20	LACE	31189993	–	C	0.949	0.339	0.948	0.961	0.946	0.078	9.14	0.229	9.32	0.096	5.98	0.073	10.22	0.058	11.58
23	rs4791868	17	LACE	9698244	–	G	0.157	0.895	0.341	0.125	0.429	0.111	9.25	0.225	9.54	0.032	6.01	0.104	10.32	0.004	11.58
24	rs7630522	3	LACE	107153088	–	T	0.703	0.199	0.914	0.938	0.679	0.013	9.26	0.219	9.76	0.143	6.15	0.106	10.43	0.006	11.59
25	rs10735825	12	LACE	50768339	FAM186A	C	0.895	0.08	0.58	0.537	0.732	0.183	9.44	0.210	9.97	0.021	6.17	0.001	10.43	0.048	11.64
26	rs730570	14	34plex	101142890	–	G	0.765	0.161	0.802	0.555	0.982	0.048	9.49	0.201	10.17	0.085	6.26	0.000	10.43	0.174	11.81
27	rs930072	5	AFR-EUR admix SNPs	36666071	SLC1A3	C	0.962	0.136	0.642	0.438	0.911	0.219	9.71	0.198	10.37	0.022	6.28	0.002	10.43	0.117	11.93
28	rs794672	6	LACE	95458317	–	G	0.057	0.917	0.432	0.741	0.357	0.246	9.96	0.197	10.57	0.020	6.30	0.017	10.45	0.025	11.95
29	rs634392	18	Genplex	70216045	–	T	0.797	0.066	0.696	0.102	0.089	0.127	10.08	0.195	10.76	0.090	6.39	0.071	10.52	0.074	12.03
30	rs4787040	16	LACE	7560980	RBFOX1	A	0.957	0.289	0.914	0.519	0.732	0.119	10.20	0.193	10.95	0.102	6.49	0.005	10.52	0.005	12.03
31	rs7307862	12	Genplex	112437514	–	C	0.576	0.899	0.142	0.426	0.839	0.000	10.20	0.172	11.13	0.209	6.70	0.010	10.53	0.046	12.08
32	rs862500	3	LACE	64272649	–	T	0.862	0.196	0.836	0.944	0.232	0.099	10.30	0.166	11.29	0.106	6.81	0.159	10.69	0.033	12.11
33	rs2503770	6	LACE	110266415	–	T	0.911	0.266	0.904	0.929	0.786	0.109	10.41	0.146	11.44	0.131	6.94	0.059	10.75	0.170	12.28
1	rs3827760	2	Genplex	109513601	EDAR	T	1	0.986	0.117	0.111	0.944	0.164	10.57	0.214	11.65	0.351	7.29	0.199	10.95	0.072	12.35
2	rs6437783	3	Genplex	108172817	MYH15	C	0.254	0.159	0.997	0.891	0.589	0.047	10.62	0.170	11.82	0.348	7.64	0.110	11.06	0.005	12.36
3	rs17822931	16	Genplex	48258198	ABCC11	C	0.997	0.854	0.096	0.615	0.875	0.180	10.80	0.077	11.90	0.346	7.99	0.000	11.06	0.060	12.42
4	rs1229984	4	Kiddlab	100239319	ADH1B	A	0	0.021	0.733	0	0.071	0.105	10.91	0.116	12.01	0.330	8.32	0.080	11.14	0.028	12.45
5	rs12594144	15	USC CCAS panel	64161351	–	C	1	0.855	0.101	0.778	0.393	0.207	11.11	0.102	12.12	0.290	8.61	0.112	11.25	0.000	12.45
6	rs4935501	10	USC CCAS panel	55935850	PCDH15	C	0.995	0.878	0.129	0.536	0.286	0.179	11.29	0.102	12.22	0.287	8.89	0.009	11.26	0.065	12.51
7	rs2180052	6	Novel	170589989	–	C	0.868	0.834	0.11	0.742	0.232	0.048	11.34	0.088	12.31	0.271	9.16	0.014	11.27	0.074	12.58
8	rs4918664	10	Kiddlab	94921065	–	A	0.989	0.874	0.131	0.141	0.821	0.175	11.52	0.106	12.41	0.271	9.44	0.146	11.42	0.027	12.61
9	rs366178	8	LACE	8771154	–	G	0.854	0.725	0.142	0.016	0.357	0.102	11.62	0.064	12.48	0.157	9.59	0.211	11.63	0.013	12.62
10	rs434504	1	DNAprint	4815477	AJAP1	G	0.168	0.2	0.902	0.481	0.268	0.063	11.68	0.081	12.56	0.256	9.85	0.007	11.64	0.016	12.64
11	rs8104441	19	Novel	51441807	–	T	0.83	0.833	0.129	0.875	0.179	0.048	11.73	0.084	12.64	0.253	10.10	0.056	11.69	0.104	12.74
12	rs1371048	2	Kersbergen	145753166	–	C	0.83	0.641	0.019	0	0.321	0.129	11.86	0.060	12.70	0.245	10.35	0.180	11.87	0.007	12.75
13	rs4657449	1	Novel	165465281	–	G	0.905	0.887	0.117	0.117	0	0.106	11.96	0.153	12.85	0.242	10.59	0.141	12.01	0.234	12.99
14	rs881929	16	34plex	31079371	ZNF668	G	0.954	0.6	0.082	0.603	0.732	0.187	12.15	0.008	12.86	0.232	10.82	0.025	12.04	0.024	13.01
15	rs10079352	5	LACE	117494640	–	G	0.038	0.451	0.972	0.984	0.429	0.266	12.42	0.018	12.88	0.229	11.05	0.166	12.20	0.010	13.02
16	rs203150	18	Novel	38037221	–	A	0.595	0.796	0.044	0.156	0	0.014	12.43	0.166	13.05	0.227	11.28	0.063	12.27	0.168	13.19
17	rs722869	14	Genplex	97277005	VRK1	C	0.895	0.901	0.168	0.259	0.429	0.078	12.51	0.137	13.18	0.224	11.50	0.102	12.37	0.019	13.21

18	rs1366220	5	LACE	153497780	–	C	0.941	0.678	0.077	0.055	0.589	0.187	12.70	0.044	13.23	0.218	11.72	0.154	12.52	0.004	13.21
19	rs499827	20	Genplex	3604447	ATRN	G	0.081	0.242	0.816	0.696	0.982	0.096	12.79	0.031	13.26	0.207	11.93	0.002	12.53	0.128	13.34
20	rs2065982	13	34plex	34864240	–	T	0.943	0.939	0.245	0.103	0.929	0.089	12.88	0.136	13.39	0.205	12.13	0.211	12.74	0.057	13.40
21	rs7246968	19	Novel	12651854	ZNF564	A	0.265	0.232	0.879	0.641	0.056	0.031	12.91	0.075	13.47	0.203	12.33	0.019	12.76	0.122	13.52
22	rs12435594	14	LACE	69555866	DCAF5	G	0.957	0.672	0.089	0.107	0.250	0.207	13.12	0.041	13.51	0.202	12.54	0.097	12.85	0.036	13.55
23	rs2851060	4	LACE	101953322	PPP3CA	C	0.097	0.224	0.871	0.946	0.411	0.119	13.24	0.076	13.59	0.200	12.74	0.181	13.04	0.001	13.55
24	rs17544484	13	Novel	88369434	–	C	0.968	0.775	0.145	0.25	0	0.179	13.42	0.067	13.65	0.195	12.93	0.059	13.09	0.225	13.78
25	rs3782973	13	Genplex	95102697	DCT	C	0.035	0.184	0.762	0.75	0.321	0.134	13.55	0.056	13.71	0.167	13.10	0.071	13.17	0.001	13.78
26	rs67302	16	DNAprint	65821799	–	T	0.859	0.511	0.066	0.219	0.411	0.163	13.71	0.011	13.72	0.164	13.26	0.027	13.19	0.000	13.78
27	rs868767	3	LACE	141377330	–	A	0.981	0.891	0.269	0.164	0.446	0.143	13.86	0.099	13.82	0.160	13.42	0.221	13.41	0.023	13.80
28	rs2833250	21	Genplex	32439270	–	T	0.392	0.046	0.699	0	0.589	0.004	13.86	0.152	13.97	0.156	13.58	0.122	13.54	0.038	13.84
29	rs1834619	2	Kiddlab	17901485	SMC6	G	1	0.929	0.304	0.047	0.286	0.158	14.02	0.122	14.09	0.152	13.73	0.274	13.81	0.082	13.92
30	rs7911953	10	LACE	61791039	ANK3	G	0.135	0.089	0.713	0.875	0.196	0.045	14.06	0.116	14.21	0.149	13.88	0.180	13.99	0.015	13.94
31	rs2585897	13	LACE	21398979	XPO4	G	0.87	0.829	0.206	0.047	0.089	0.083	14.15	0.101	14.31	0.146	14.03	0.204	14.19	0.150	14.09
1	rs433342	8	USC Tri-allic SNP	17747876	FGL1	AG	0.40/0.17	0.30/0.66	0.56/0.44	0.87/0.13	0.19/0.81	0.163	14.31	0.049	14.36	0.048	14.07	0.169	14.36	0.102	14.19
2	rs2069945	20	USC Tri-allic SNP	33761837	PROCR	CG	0.11/0.71	0.46/0.48	0.65/0.27	0.04/0.54	0.77/0.23	0.126	14.44	0.006	14.36	0.039	14.11	0.104	14.47	0.191	14.38
3	rs4540055	4	34plex	38803255	TLR1	AC	0.07/0.40	0.68/0.25	0.37/0.51	0.54/0.21	0.60/0.40	0.183	14.62	0.062	14.43	0.021	14.13	0.104	14.57	0.022	14.40
4	rs5030240	11	34plex	32424389	WT1	CA	0.39/.0.32	0.69/0.12	0.28/0.05	0.09/0.28	0.25/0	0.059	14.68	0.089	14.52	0.059	14.19	0.067	14.64	0.078	14.48
5	rs17287498	10	USC Tri-allic SNP	54530788	MBL2	GA	0.48/0.49	0.59/0.19	0.77/0.10	0.96/0.02	0.98/0.01	0.096	14.77	0.021	14.54	0.017	14.21	0.145	14.78	0.075	14.55
6	rs2184030	1	USC Tri-allic SNP	206667441	–	CG	0.59/0.19	0.48/0.51	0.58/0.41	0.48/0.52	0.72/0.02	0.072	14.85	0.024	14.56	0.022	14.23	0.042	14.83	0.156	14.71

Component SNPs are ranked in descending order of informativeness (defined by individual population-specific divergence = PSD) in population groups (PG): Oceanian (28); American (22); African (8); European (33) and East Asian (31). Six tri-allelic SNPs are listed at the end with allele frequencies of 1000 Genomes listed alleles. C: chromosome; RA: reference allele; CD: cumulative PSD (derived by addition of individual PSDs). SNP positions are from genome build 37.1 (GRCh37).

ancestry informativeness. Top components are: rs2814778; rs16891982; rs1426654; rs3827760; rs2789823; rs1871534; rs1369290; rs6875659; rs310644; rs9908046. Only the last three loci are part of the Illumina 650 K WGS panel often used for AIM-SNP selection.

Supplementary Fig. S3 charts the inter-marker Mb spans on each chromosome, indicating a median 14.8 Mb distance amongst syntenic Global AIM-SNPs, while highlighting four pairs with below average close physical linkage (including two transgressions of the 1 Mb filter). HapMap data indicated absence of correlated allele frequencies across 500 kb distances (maximum possible in Haploview) within each span. Therefore, these SNP pairs lack allelic correlation and unless very recently admixed (e.g. parentally), correlations between ancestry-informative alleles in individuals with co-ancestry will erode rapidly [33].

3.2. Balancing population-specific Divergence levels and subsets of the full Global AIM-SNP set

The five sets of SNPs listed in Table 1 and Supplementary Table S2A that focus on each global population group comprise: 33 EUR-informative; 31 E ASN-informative; 28 OCE-informative; 22 AME-informative and 8 AFR-informative markers, with a further six tri-allelic SNPs added. The 122-SNP set combined in these proportions reached cumulative PSD values of EUR = 14.31, E ASN = 14.03, OCE = 14.09, AME = 14.19, AFR = 14.15 (a standard deviation of $\pm\sim$0.1). This very balanced distribution of PSD values in each of the five groups is maintained adding the six tri-allelic markers, giving cumulative PSD values of EUR = 14.56, E ASN = 14.23, OCE = 14.71, AME = 14.82, AFR = 14.84 $\pm\sim$0.22). Therefore this SNP set has near-identical power to differentiate all five population groups, allows balanced analysis of each of the ten pairwise group comparisons and can be used to assess patterns of admixture without exaggerating the contribution of any one group due to an excess of informative markers for that group compared to the others. Fig. 4A charts the progression toward a point of PSD convergence, combining 122 + 6 component SNPs by carefully adjusted stepwise accumulation of divergence across five groups.

Reaching a point of PSD convergence for 128 SNPs, we next sought to reduce component numbers to subsets of approximately 90, 60 and 30 SNPs or less, while preserving PSD balance. The medium-scale subsets consisted of 88 SNPs with PSD values of EUR = 10.58, E ASN = 10.52, OCE = 10.47, AME = 10.36, AFR = 10.35 ($\pm\sim$0.09) and 55 SNPs with PSD values of EUR = 6.99, E ASN = 6.98, OCE = 6.81, AME = 7.07, AFR = 7.12 ($\pm\sim$0.1), listed in Supplementary Table S3. Construction of smaller scale subsets from the most divergent SNPs, examined how low the number of AIMs can be reduced while preserving both balance and classification success (using *Snipper* cross-validation). The three smallest subsets are summarized in Supplementary Fig. S4, which shows 12 SNPs reach a three group PSD of 3 and 100% classification success, extending this subset to 20 SNPs adds 100% classification success for Americans and extension to 28 provides reasonable PSD balance across four groups, converging on 4.7–5.0.

3.3. The Sequenom iPLEX Global genotyping tests

Amongst the five optimized iPLEX tests, 12/128 SNPs were not incorporated, a 90.6% assay conversion rate. Supplementary Fig. S1B outlines five alternatives from the same divergent chromosome segments with four showing near-identical allele frequency distributions. The failed AFR-informative SNP rs1871534 could be replaced with equivalent markers rs4598087 or rs7778058 and failed EUR-informative rs17287498 with rs2855557. Nine additional SNPs are in the iPLEX tests for allele frequency validation

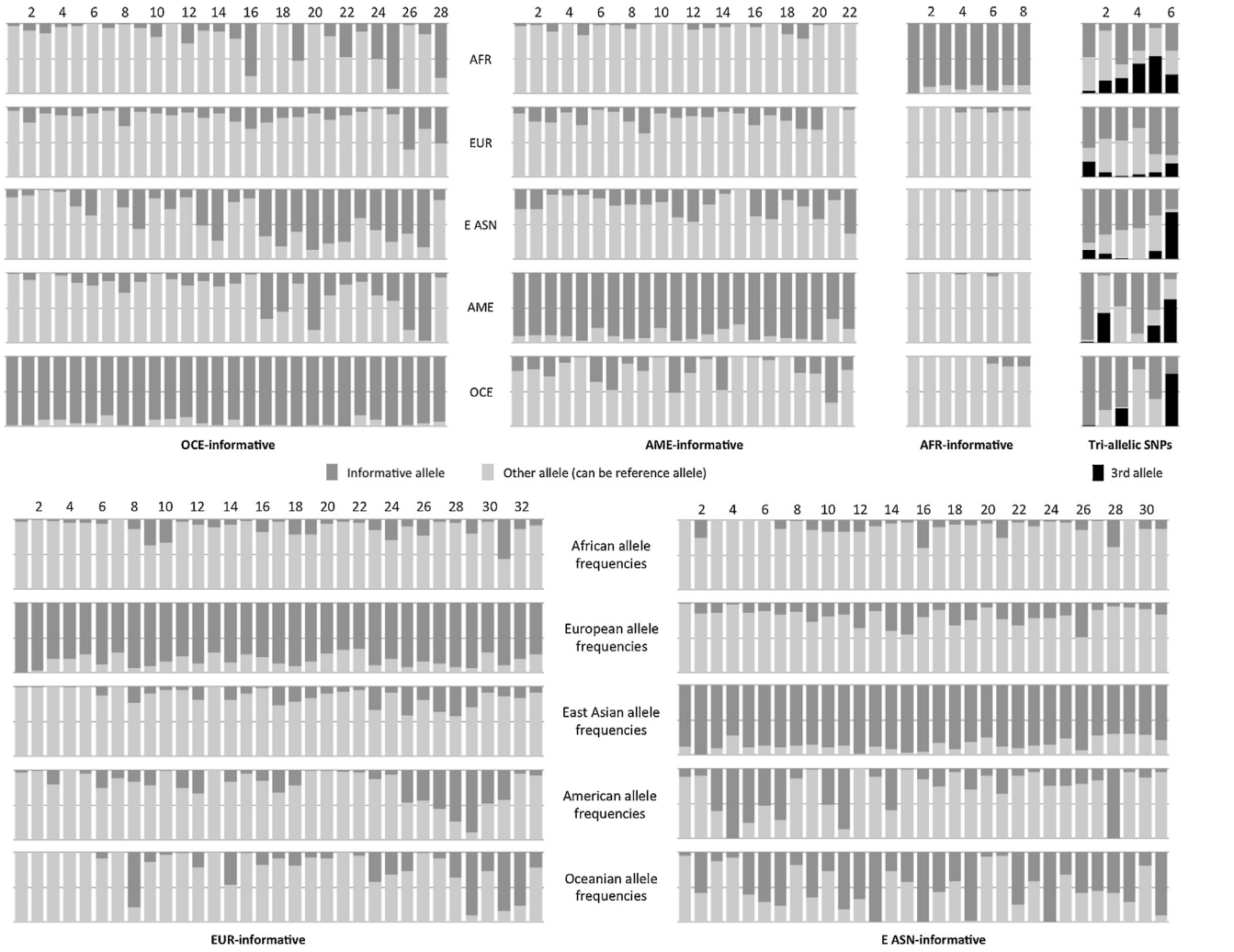

Fig. 2. Allele frequency distributions of the Global AIM-SNP set arranged into the five groups they best differentiate plus tri-allelic SNPs. Column numbers identify individual SNPs ranked by population-group informativeness in Table 1. Mid-gray bars in tri-allelic SNPs denote reference allele, light gray and black bars 2nd and 3rd alleles respectively.

purposes (including rs4598087, rs7778058, rs2855557) so all three alternatives are already included.

Sequencing confirmed iPLEX allele calls made from imbalanced peaks in rs2789823, rs8137373, rs1557553 and rs2471552, with no clustering SNP sites observed. Therefore allele peak balance parameters were adjusted in the R scripts to accommodate the peak height skews observed. One partially silent allele was identified for rs12913832, with sequencing revealing an A to G variant in position 3 of the reverse PCR primer (data not shown). This variant likely reduces amplification efficiency of the rs12913832-A allele and was not previously reported in dbSNP, 1000 Genomes, or amongst users of the 34plex ancestry test and Irisplex/HIrisplex forensic pigmentation-predictive tests that all include this SNP [3,4,7].

Initial forensic sensitivity assessments of the five iPLEX tests are summarized in Supplementary File S1C. Optimum template DNA was gauged to be $\geq$1.25 ng for Global2, 3 and 4, but somewhat higher for Global1 and 5 at $\geq$5 ng. This data suggests these MALDI-TOF tests will provide a viable forensic technique as well as a simple system for generating extended population data to further explore global variability in the Global AIM-SNPs.

3.4. Analysis of admixed populations with the Global AIM-SNP set

PCA and STRUCTURE analyses of reference and admixed populations are summarized in Fig. 5. The uppermost plot of Fig. 5A shows STRUCTURE results for an optimum K:5 genetic clusters (indicated on the likelihood-of-K chart right): differentiating five reference groups and four admixed populations. Almost identical patterns are seen in the paired plots below for an optimum K:3 genetic clusters comparing the same admixed samples (but with three reference groups) using 128 Global SNPs and 350 LACE SNPs. While visual inspection of cluster plots suggests a close match between the co-ancestry patterns seen in ASW, CLM, MXL and PUR individuals for 128 vs. 350 AIM-SNPs, individual co-ancestry proportion estimates that underlie the cluster plots provide a more precise comparison. Supplementary Tables S4 lists 187 individual co-ancestry proportions and their averages in the four admixed populations for AFR-EUR-AME co-ancestry components. Global SNPs gave average proportions for ASW of: AFR = 73.7%-EUR = 21.6%-AME = 4.7%; CLM = 9%-67.3%-23.7%; MXL = 8.1%-46.7%-45.2%; PUR = 12.1%-74.7%-13.2%; while LACE gave: ASW of: AFR = 75.3%-EUR = 20.8%-AME = 3.9%; CLM = 9.3%-66.8%-23.9%; MXL = 4.5%-49.0%-46.5%; PUR = 12.9%-75.0%-12.1%. These comparisons show minimal differences in co-ancestry proportion estimates between each SNP set of about 1%, reaching a maximum 3.6% difference for the AFR component estimates in MXL. The very close match between SNP sets is also shown by the mean and quartile differences in individual co-ancestry proportion estimates listed and charted in Supplementary Table S4. The great majority of estimates differ by much less than

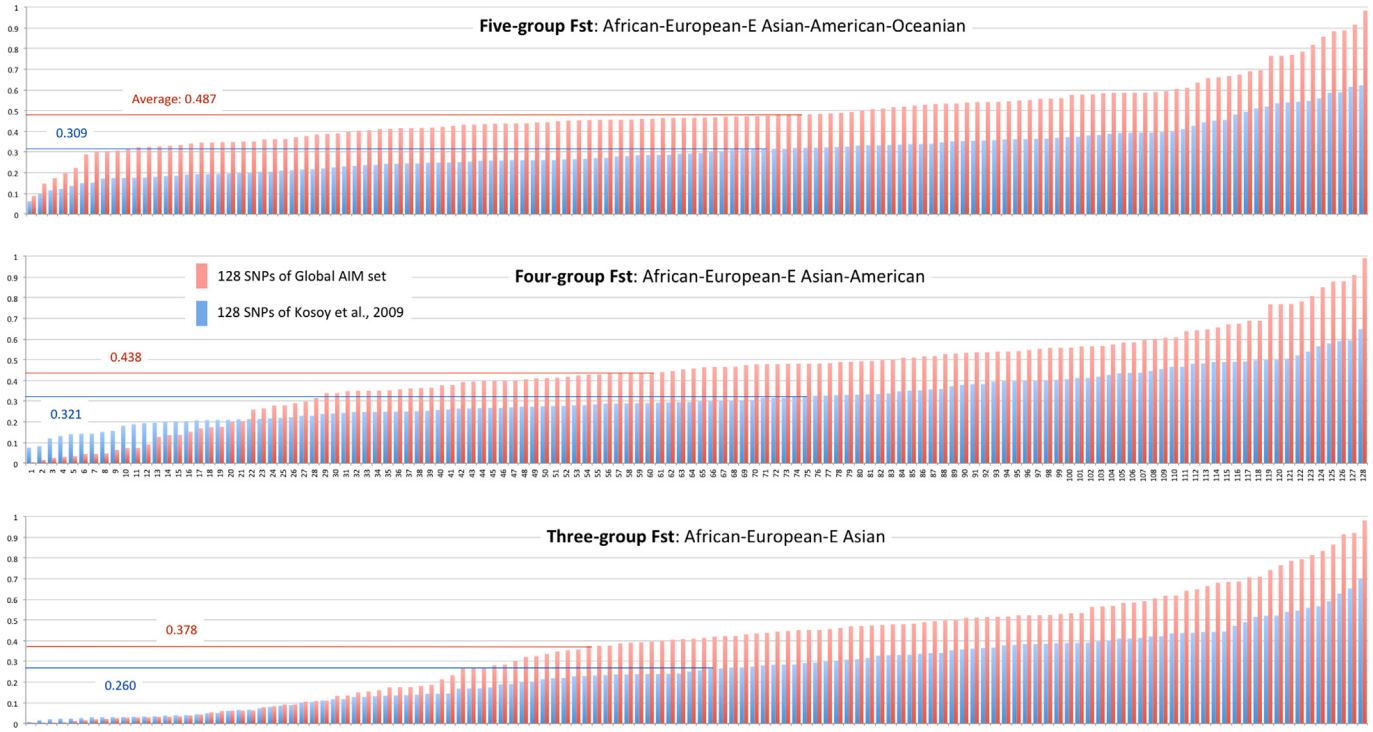

Fig. 3. Ranked Fst values for three, four and five group comparisons analyzed with 128 Global AIM-SNPs or with the closest equivalent panel: 128 SNPs of the AFR-EUR-AME-informative ancestry panel of Kosoy et al. [20].

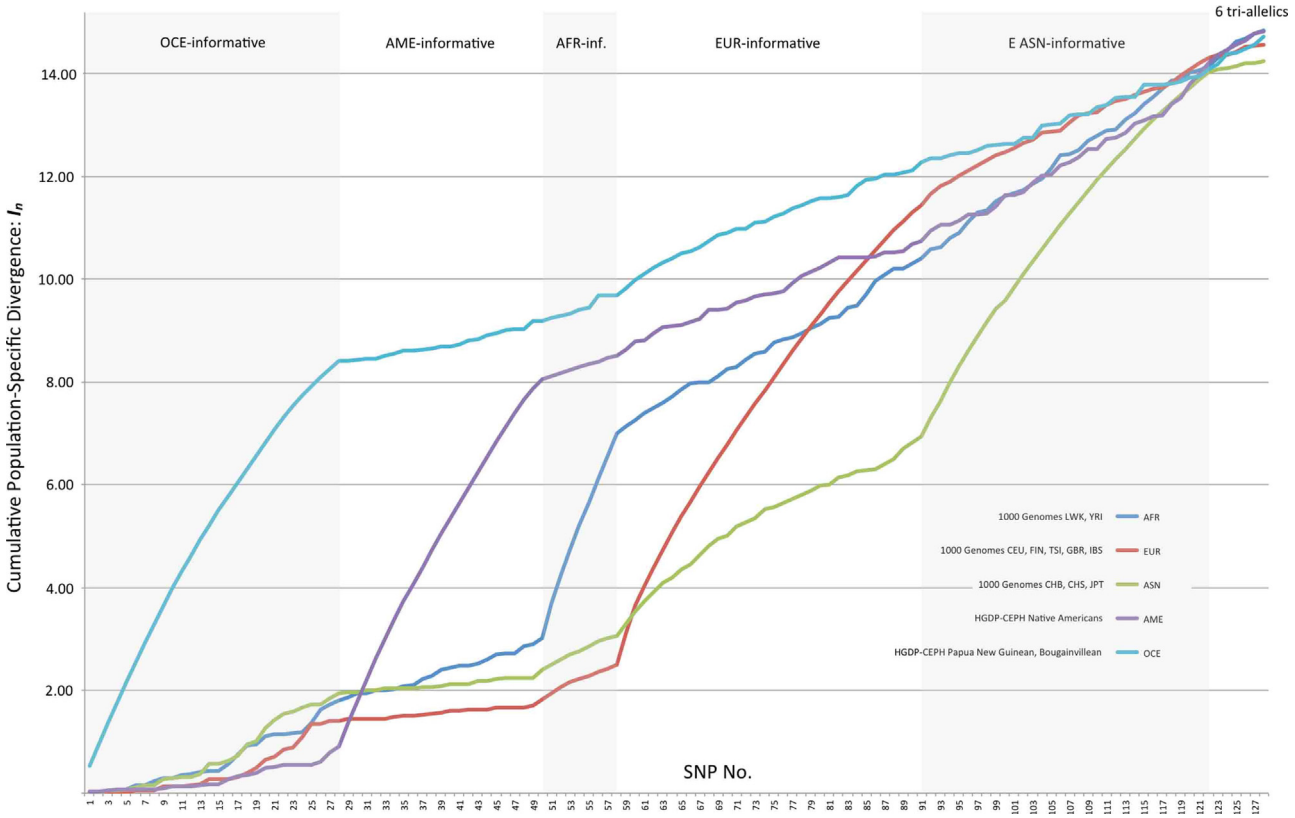

Fig. 4. Convergence of five different cumulative population-specific Divergence (PSD) values in the 128 Global AIM-SNP set with components ordered by OCE-informative, best to worst differentiators, then: AME-; AFR-; EUR- and E ASN-informative. 122 SNPs reach a PSD convergence value of ~14 then tri-allelic SNPs introduce some dispersion of PSD balance, notably in E ASNs. Marker order is the same as the AIM-SNP list in Supplementary Table S2A.

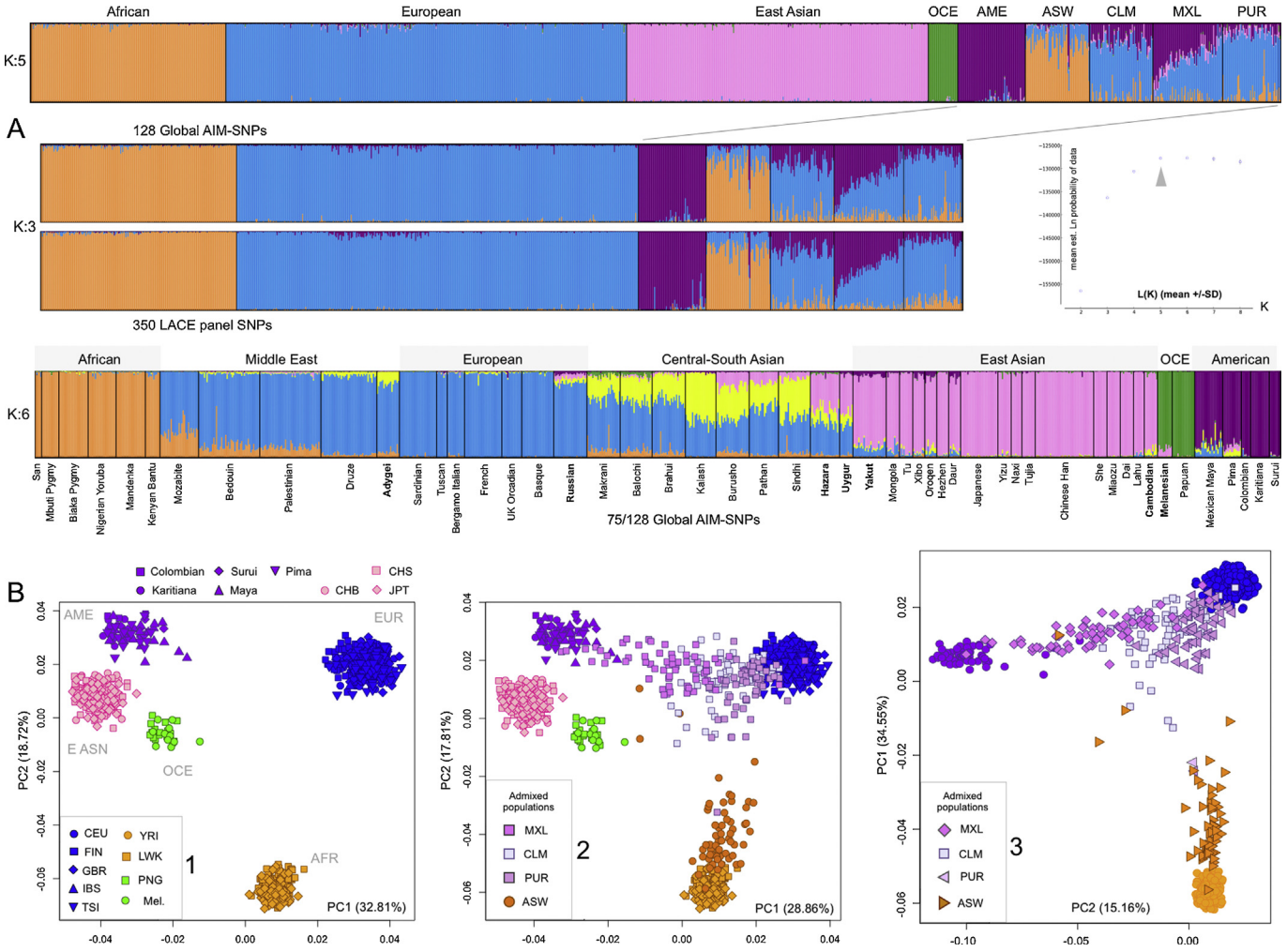

Fig. 5. Population analyses of 1000 Genomes and HGDP-CEPH SNP data. (A) STRUCTURE cluster plots analyzing, top to bottom: five reference and four admixed populations, optimum K:5 (indicated in L(K) chart right); three reference (AFR-EUR-AME) and admixed populations, optimum K:3 with 128 Global vs. 350 LACE SNPs; all HGDP-CEPH populations, optimum K:6 with a subset of 75/128 Global SNPs. Bold population labels below indicate seven outlier populations that show strong similarities to cluster plot patterns obtained by Li et al. (Fig. 1 of [17]) analyzing the same samples at K:7 with 650,000 SNPs. The order of populations in each group is the same as Li's cluster plot to allow easy comparison. (B1) PCA 2-PC plot with 128 SNPs analyzing the same population combinations as the top STRUCTURE results. Colors matched to each reference group cluster. (B2) PCA adding four 1000 Genomes admixed populations. (B3) PCA of the second STRUCTURE plot population combinations, comprising 128 SNPs and three reference groups.

5% between each SNP set, reaching a maximum 8% difference for the 75% quartile in MXL, AFR component estimates. Therefore the 128 SNPs we have collected provide a very close match, in terms of admixture detection capabilities, to a powerful AIM-SNP set almost three-fold larger and centered in its selection on three of the five population groups the Global set differentiates.

Fig. 5B shows three PCA plots that link to the STRUCTURE cluster plots above. PCA with 128 Global SNPs of reference populations alone (Fig. 5B1) estimated the first three principal components (PC) accounted for PC1 = 32.8%, PC2 = 18.7% and PC3 = 6.4% of the total variation. In the case of the middle and right-hand PCA plots (Figs. 5B2-3), visual inspection is the only means to assess how well the co-ancestry analysis capabilities of both SNP sets match. Cluster patterns generally appear well matched with very similar outlier positions for three mid-cluster ASW and one PUR. These plots suggest the smaller 128 SNP Global set would give consistent positioning of admixed individuals on PC1-PC2 plots compared with the larger LACE set, in agreement with the findings from STRUCTURE analyses. The PC1-PC2 and PC2-PC3 plots from the left-hand PCA analysis of Fig. 5B are shown in Supplementary Fig. S5. Individual PCAs (and PC1-PC2 estimates) for the four admixed 1000 Genomes study populations combined in Fig. 5B2 are also included.

3.5. Assessment of within-population group heterogeneity of selected AIM-SNPs

Analysis of within-population, within-group and between-group SNP allele frequency heterogeneity was assessed with pairwise Fst values and average pairwise genotype differences generated by Arlequin. Analyses compared the 1000 Genomes-CEPH populations used for AIM selection with additional CEPH AFR, EUR and E ASN populations or with four novel divergent populations, and are summarized in Fig. 6. Additionally, STRUCTURE analyses were made of the complete HGDP-CEPH panel, restricted to the 75/128 SNPs in the Stanford data. The resulting plot of the optimum K:6 clusters is shown in the lower graphic of Fig. 5A. Genetic cluster patterns for novel AFR, EUR and E ASN populations are mainly the same as their 1000 Genomes counterparts. The sixth, yellow cluster is confined to Central-South Asians and Caucasian Adygei (the most closely sited European population). However, Central-South Asian and Middle East populations

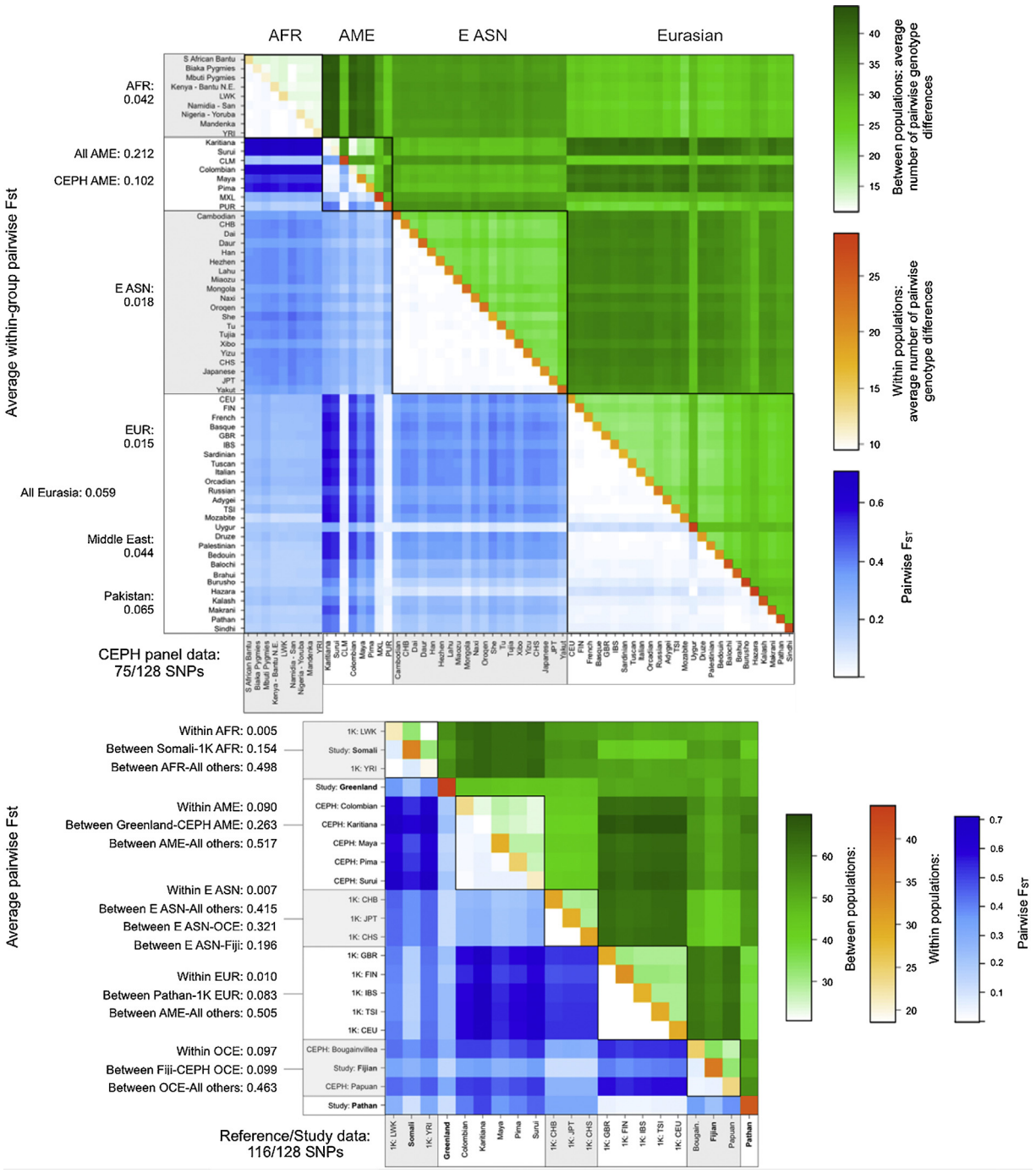

Fig. 6. Arlequin population divergence analyses of pairwise Fst (blue squares), pairwise between-population genotype differences (green) and pairwise within-population differences (orange). Upper chart plots comparisons of HGDP-CEPH populations with equivalent 1000 Genomes populations, restricted to 75/128 SNPs. Lower chart shows comparisons of four additional study populations likely to be divergent from 1000 Genomes/CEPH populations used for AIM selection. Upper chart values, listed left, show within-group Fst averages (lower half of large diagonal boxes) indicating very low within-group divergence is maintained adding HGDP-CEPH populations to those used for selection. Raised Fst is evident within the American group, particularly amongst admixed 1000 Genomes CLM, MXL, PUR. Lower chart Fst values on the left indicate a uniformly high ratio of between-group to within-group divergence. However, divergence between Greenland and other AME populations is very high, while there is markedly reduced divergence between East Asians and Fijians – values likely reflecting different demographic histories between Arctic vs. Meso-S American populations and Near vs. Far Oceanian populations respectively.

do not produce distinctive clusters since Global AIM selection did not target differentiation of Europeans from these Eurasian sub-groups. A comparison of the STRUCTURE cluster plot with the comprehensive analysis of the same samples by Li et al. that detected seven clusters (Fig. 1 in [17]), indicates some closely matching patterns for many populations, notably: Adygei, Russian, Hazara, Uygur, Yakut, Cambodian and Melanesian. Overall, despite the constraint of analyzing only 60% of SNP data, STRUCTURE analysis of 30 new AFR, EUR, E ASN populations does not reveal significant allele frequency heterogeneity. Similarly, Arlequin analysis shows low within-group divergences for the same groups (first, third and fourth large diagonal boxes, upper plot of Fig. 6).

A lack of heterogeneity amongst the selected AIMs can be attributed to the loci as much as the populations we added to

extended analyses, i.e. in SNPs at or near to fixation there is limited scope for drift to strongly influence very low frequency alleles. In contrast, a review of certain Arlequin pairwise comparisons in Fig. 6 reveals admixture or divergent population origins can create higher levels of heterogeneity that may affect how well the AIMs we selected differentiate all populations worldwide. 1000 Genomes CLM, MXL and PUR patterns indicate substantial divergence can occur from admixture. All three show less divergence from EUR populations than AME, as well as raised within-population heterogeneity.

Lastly, differences in the origins and demographics of populations placed in the same group due to geographic proximity can create the largest levels of within-group heterogeneity. Greenlanders produced distinctive patterns of pairwise Fst, shown in the lower plot of Fig. 6, suggesting they are different to all the populations we used for AIM selection. While EUR admixture may contribute some divergence, most of the heterogeneity seen is more likely to be due to the unique peopling of Greenland in two very recent migratory waves: from Siberia by Paleo-Eskimos 4500 years ago and from Beringia by Inuit 1000 years ago [34] – in stark contrast to the separate, earlier colonization of the rest of America. Some heterogeneity is also seen in Fijians, not in increased divergence within OCE, but in much reduced divergence with E ASN. The observation of divergence between Near and Far Oceania is, again, likely to reflect differences in the origins of the migrant populations that first colonized two distinct parts of the same continent (reviewed in Chapter 13 of [16]) as well as recent migration from e.g. South Asians. While it can be argued that strong divergence within a group will compromise the ability of the AIMs selected to fully differentiate populations, when we construct *Snipper* training sets of both Greenlanders and Fijians, in place of AME and OCE data or in addition to the five groups (as provided in Supplementary File S1), all samples are assigned correctly to their population of origin.

4. Discussion

This study shows the current extensive human genome variation catalogs can be easily accessed and their allele frequency data used to select highly differentiating ancestry informative SNPs. We were able to build sets with a range of sizes that meet the statistical power demands of forensic analysis, while focusing on the key characteristic of population differentiation balance. Although prompted by the previous study of Galanter, that addressed AFR-EUR-AME populations [22], for all but the smallest subsets, we successfully kept five population group comparisons equally well differentiated and did not confine SNP informativeness to three or four global groups. Up until now, many AIM-SNP selections have ignored balanced differentiation of East Asians in favor of the closely related Native American group. Such approaches preclude unbiased and globally applicable analysis of a forensic sample of unknown geographic origin, despite East Asian admixture being rare in The Americas, and despite Oceanian ancestry representing a minor proportion of the worldwide demographic landscape outside of Australia, New Zealand and Hawaii.

The SPSmart genome browsers [13,14] proved to be particularly valuable aids to building comprehensive forensic AIM-SNP sets in three respects. Firstly, they provide the only viable means to interrogate hundreds of markers with specific population properties in the same query, whereas 1000 Genomes, HapMap and HGDP-CEPH data gateways only allow SNPs to be screened locus-by-locus. Secondly, the option to identify groups of SNPs with identical allele frequency patterns, and therefore divergence [15], is straightforward in SPSmart by defining the chromosome segment to be queried, thus providing alternative markers for many first-strike SNPs that fail in multiplex or NGS analysis due to poor context sequence. Thirdly, SPSmart enables a very simple system for downloading genotype data from selected population groupings for porting into *Snipper* to assess patterns of divergence amongst markers and population pairings chosen.

With our study's primary focus on equilibrating the differentiation power evenly amongst five population groups, it is important to assess how well the Global AIM-SNP set can gauge admixture between them and whether forensic analysis has the right tools for the task. Our assessments of both of these aspects remain somewhat limited by the availability of data, being confined to four populations from The Americas. The imminent data release of 1000 Genomes Peruvians from Lima (PEL) may reveal East Asian admixture components, while the recent detailed study of American populations by Reich et al. also provides useful SNP data [35]. Analysis of Fijians with 116/128 SNPs suggests that divergence between Near and Far Oceania is detectable but small, although recent migration and integration of South Asians to Fiji is an additional factor that should be considered in any interpretation of ancestry patterns. Furthermore, the distinction between populations with admixture and those on continental margins where absence of geographic barriers has led to increased gene-flow, is a complicating factor that requires caution and makes population analyses challenging when moving beyond unadmixed 'continental' populations such as the reference data used in this study. For example, it is not straightforward to distinguish allele frequency patterns found in North Africans from those found in admixed US African-Americans (see Fig. 4 of [28]). STRUCTURE provides the most commonly applied system for estimating co-ancestry components in admixed individuals by accurately defining the optimum number of genetic clusters in the data [18,26,30]. However, there are widely held doubts about the validity of using mid-continental reference populations and not taking sufficient account of the clinal variation characteristic of so much of worldwide demographic structure [36,37]. A simpler and more easily interpreted approach to the complex task of distinguishing patterns of admixed SNP variation from those of unadmixed ancestral populations, is to superimpose individuals with unknown ancestry alongside reference population data on distance matrices generated by multi-dimensional scaling systems such as PCA. In forensic analysis it is beneficial to have a flexible system of ancestry inference that can easily handle single SNP profiles without using lengthy and PC-intensive calculations (typified by STRUCTURE). Therefore we have improved forensic ancestry analysis using real-time tools in two ways. Firstly, by adding a co-ancestry component estimator in *Snipper* based on the ratio of each likelihood to the sum of all likelihoods (the number of ancestries defined by the training set data uploaded). Secondly, by developing scripts that position de-novo AIM-SNP profiles on top of reference data PCA plots (such as Fig. 5B) using the first and second principal component percentages to define the profile's plot axes [http://mathgene.usc.es/snipper/analysismultipleprofiles.html]. In this way, visualization of complex genetic data in simple 2-D space can bring a more intuitive approach to ancestry inference, aided by this study's advances in assembling some of the most informative AIM-SNPs for forensic use.

Acknowledgements

This work was funded by the EUROFORGEN Node of Excellence (Grant Agreement No. 285487). Studies leading to the reported results were financially supported by the Austrian Science Fund (FWF, P22880-B12). CS is supported by funding awarded by the Portuguese Foundation for Science and Technology (FCT) and co-financed by the European Social Fund (Human Potential Thematic Operational Program SFRH/BD/75627/2010).

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:10.1016/j.fsigen.2014.02.012.

References

[1] S.B. Seo, J.L. King, D.H. Warshauer, C.P. Davis, J. Ge, B. Budowle, Single nucleotide polymorphism typing with massively parallel sequencing for human identification, Int. J. Legal Med. (2013), http://dx.doi.org/10.1007/s00414-013-0879-7.

[2] D.H. Warshauer, D. Lin, K. Hari, R. Jain, C. David, B. Larue, J.L. King, B. Budowle, STRait Razor: A length-based forensic STR allele-calling tool for use with second generation sequencing data, Forensic Sci. Int. Genet. 7 (2013) 409–417.

[3] S. Walsh, F. Liu, K.N. Ballantyne, M. van Oven, O. Lao, M. Kayser, DNA-based eye colour prediction across Europe with the IrisPlex system, Forensic Sci. Int. Genet. 5 (2011) 170–180.

[4] S. Walsh, F. Liu, A. Wollstein, L. Kovatsi, A. Ralf, A. Kosiniak-Kamysz, W. Branicki, M. Kayser, The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA, Forensic Sci. Int. Genet. 7 (2013) 98–115.

[5] Y. Ruiz, C. Phillips, A. Gomez-Tato, J. Alvarez-Dios, M. Casares de Cal, R. R. Cruz, O. Maroñas, J. Söchtig, M. Fondevila, M.J. Rodriguez-Cid, Á. Carracedo, M.V. Lareu, Further development of forensic eye color predictive tests, Forensic Sci. Int. Genet. 7 (2013) 28–40.

[6] R. Pereira, C. Phillips, N. Pinto, C. Santos, C.E.B. Santos, A. Amorim, Á. Carracedo, L. Gusmão, Straightforward inference of ancestry and admixture proportions through ancestry-informative insertion deletion multiplexing, PLoS One 7 (2012) e29684.

[7] C. Phillips, A. Salas, J.J. Sánchez, M. Fondevila, A. Gómez-Tato, J. Álvarez-Dios, M. Calaza, M. Casares de Cal, D. Ballard, M.V. Lareu, Á. Carracedo, The SNP*for*ID consortium, inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs, Forensic Sci. Int. Genet. 1 (2007) 273–280.

[8] M. Fondevila, C. Phillips, C. Santos, A. Freire Aradas, P.M. Vallone, J.M. Butler, M.V. Lareu, Á. Carracedo, Revision of the SNP*for*ID 34-plex forensic ancestry test: assay enhancements, standard reference sample genotypes and extended population studies, Forensic Sci. Int. Genet. 7 (2013) 63–74.

[9] C. Phillips, Applications of autosomal SNPs and Indels in forensic analysis, Forensic Sci. Rev. 42 (2012) 44–62.

[10] M. Fondevila, C. Phillips, N. Naveran, L. Fernandez, M. Cerezo, A. Salas, Á. Carracedo, M.V. Lareu, Case report: identification of skeletal remains using short-amplicon marker analysis of severely degraded DNA extracted from a decomposed and charred femur, Forensic Sci. Int. Genet. 2 (2008) 212–218.

[11] C. Phillips, M. Fondevila, M.V. Lareu, A 34-plex autosomal SNP single base extension assay for ancestry investigations, Methods Mol. Biol. 830 (2012) 109–126.

[12] C. Phillips, A. Freire Aradas, A.K. Kriegel, M. Fondevila, O. Bulbul, C. Santos, F. Serrulla Rech, M.D. Perez Carceles, Á. Carracedo, P.M. Schneider, M.V. Lareu, Eurasiaplex: A forensic SNP assay for differentiating European and South Asian ancestries, Forensic Sci. Int. Genet. 7 (2013) 359–366.

[13] J. Amigo, A. Salas, C. Phillips, Á. Carracedo, SPSmart: adapting population based SNP genotype databases for fast and comprehensive web access, BMC Bioinform. 9 (2008) 428.

[14] J. Amigo, A. Salas, C. Phillips, *ENGINES*: exploring single nucleotide variation in entire human genomes, BMC Bioinform. 12 (2011) 105.

[15] J. Costas, A. Salas, C. Phillips, Á. Carracedo, Human genome-wide screen of haplotype-like blocks of reduced diversity, Gene 349 (2005) 219–225.

[16] M.A. Jobling, E. Hollox, M.E. Hurles, T. Kivisild, C. Tyler-Smith, Human Evolutionary Genetics, 2nd ed., Garland Science, New York, 2014.

[17] J.Z. Li, D.M. Absher, H. Tang, A.M. Southwick, A.M. Casto, S. Ramachandran, H.M. Cann, G.S. Barsh, M. Feldman, L.L. Cavalli-Sforza, R.M. Myers, Worldwide human relationships inferred from genome-wide patterns of variation, Science 319 (2008) 1100–1104.

[18] N.A. Rosenberg, J.K. Pritchard, J.L. Weber, H.M. Cann, K.K. Kidd, L.A. Zhivotovsky, M.W. Feldman, Genetic structure of human populations, Science 298 (2002) 2381–2385.

[19] I. Halder, M. Shriver, M. Thomas, J.R. Fernandez, T. Frudakis, A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications, Hum. Mutat. 29 (2008) 648–658.

[20] R. Kosoy, R. Nassir, C. Tian, P.A. White, L.L.M. Butler, G. Silva, R. Kittles, M.E. Alarcon-Riquelme, P.K. Gregersen, J.W. Belmont, F.M. De La Vega, M.F. Seldin, Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America, Hum. Mutat. 30 (2009) 69–78.

[21] J.R. Kidd, F.R. Friedlaender, W.C. Speed, A.J. Pakstis, F.M. De La Vega, K.K. Kidd, Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples, Invest. Genet. 2 (2011) 1.

[22] J.M. Galanter, J.C. Fernandez-Lopez, C.R. Gignoux, J. Barnholtz-Sloan, C. Fernandez-Rozadilla, M. Via, A. Hidalgo-Miranda, A.V. Contreras, L. Uribe Figueroa, et al., Development of a panel of genome-wide ancestry informative markers to study admixture throughout the Americas, PLoS Genet. 8 (2012) e1002554.

[23] P. Kersbergen, K. van Duijn, A.D. Kloosterman, J.T. den Dunnen, M. Kayser, P. de Knijff, Developing a set of ancestry-sensitive DNA markers reflecting continental origins of humans, BMC Genet. 10 (2009) 69.

[24] C.M. Nievergelt, A.X. Maihofer, T. Shekhtman, O. Libiger, X. Wang, K.K. Kidd, J.R. Kidd, Inference of human continental origin and admixture proportions using a highly discriminative ancestry informative 41-SNP panel, Invest. Genet. 4 (2013) 13, (NB: [24] describes 41 of 55 SNPs listed in FROGkb: http://frog.med.yale.edu/FrogKB/

[25] C. Phillips, R. Fang, D. Ballard, M. Fondevila, C. Harrison, F. Hyland, E. Musgrave-Brown, C. Proff, E. Ramos-Luis, B. Sobrino, M. Furtado, D. Syndercombe Court, Á. Carracedo, P.M. Schneider, The SNP*for*ID consortium, Evaluation of the Genplex SNP typing system and a 49-plex forensic marker panel, Forensic Sci. Int. Genet. 1 (2007) 180–185.

[26] C. Borel, F. Cheung, H. Stewart, D.A. Koolen, C. Phillips, N.S. Thomas, P.A. Jacobs, S. Eliez, A.J. Sharp, Evaluation of PRDM9 variation as a risk factor for recurrent genomic disorders and chromosomal non-disjunction, Hum. Genet. 131 (2012) 1519–1524.

[27] N.A. Rosenberg, L.M. Li, R. Ward, J.K. Pritchard, Informativeness of genetic markers for inference of ancestry, Am. J. Hum. Genet. 73 (2003) 1402–1422.

[28] C. Phillips, Ancestry informative markers, in: J.A. Siegel, P.J. Saukko (Eds.), 2nd ed., Encyclopedia of Forensic Sciences, vol. 1, Academic Press, 2013, pp. 323–331.

[29] T. Bersaglieri, P.C. Sabeti, N. Patterson, T. Vanderploeg, S.F. Schaffner, J.A. Drake, M. Rhodes, D.E. Reich, J.N. Hirschhorn, Genetic signatures of strong recent positive selection at the lactase gene, Am. J. Hum. Genet. 74 (2004) 1111–1120.

[30] P. P. Johansen, J.D. Andersen, C. Børsting, N. Morling, Evaluation of the iPLEX® Sample ID Plus Panel designed for the Sequenom MassARRAY® system. A SNP typing assay developed for human identification and sample tracking based on the SNP*for*ID panel, Forensic Sci. Int. Genet. 7 (2013) 482–487.

[31] L. Porras-Hurtado, Y. Ruiz, C. Santos, C. Phillips, Á. Carracedo, M.V. Lareu, An overview of STRUCTURE: applications, parameter settings, and supporting software, Front. Genet. 4 (2013) 98.

[32] J.R. Gonzalez, L. Armengol, X. Sole, E. Guino, J.M. Mercader, X. Estivill, V. Moreno, SNPassoc: an R package to perform whole genome association studies, Bioinformatics 23 (2007) 644–645.

[33] P. Gill, C. Phillips, C. McGovern, J.A. Bright, J. Buckleton, An evaluation of potential linkage disequilibrium between the STRs vWA and D12S391 with implications in criminal casework, Forensic Sci. Int. Genet. 6 (2012) 477–486.

[34] V. Colonna, L. Pagani, Y. Xue, C. Tyler-Smith, A world in a grain of sand: human history from genetic data, Genome Biol. 12 (2011) 234.

[35] D. Reich, N. Patterson, D. Campbell, A. Tandon, S. Mazieres, N. Ray, M.V. Parra, W. Rojas, C. Duque, et al., Reconstructing Native American population history, Nature 488 (2012) 370–374.

[36] D. Serre, S. Pääbo, Evidence for gradients of human genetic diversity within and among continents, Genome Res. 14 (2004) 1679–1685.

[37] N.A. Rosenberg, S. Mahajan, S. Ramachandran, C. Zhao, J.K. Pritchard, M.W. Feldman, Clines, clusters, and the effect of study design on the inference of human population structure, PLoS Genet. 1 (6) (2005) e70.